CEIG - Spanish Computer Graphics Conference (2022)
J. Posada and A. Serrano (Editors)

DriveRNN: Predicting Drivers' Attention with Deep Recurrent Networks

Blanca Lasheras-Hernandez, Belen Masia, and Daniel Martin

Universidad de Zaragoza, I3A

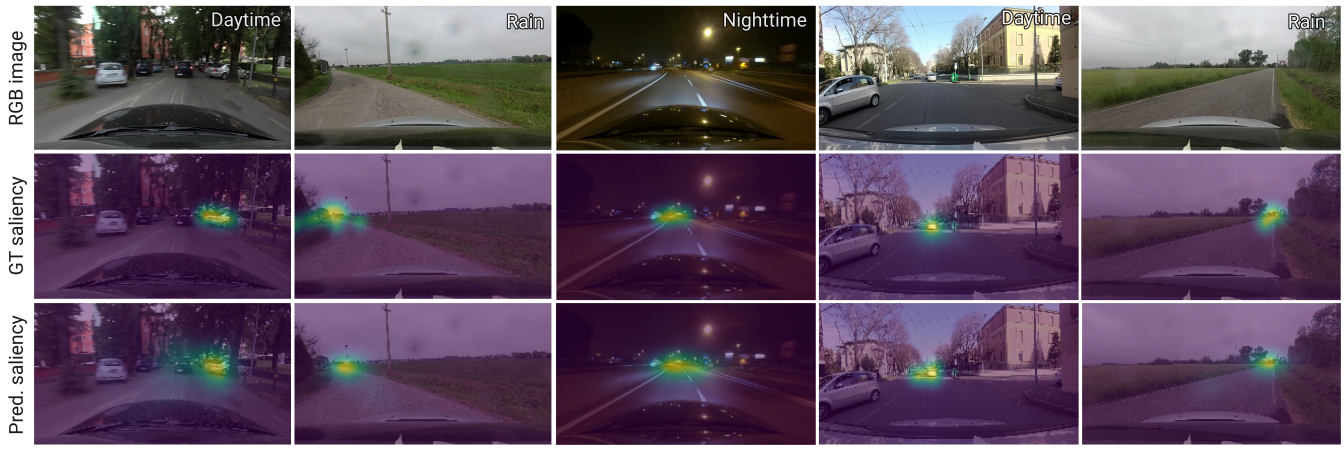

Figure 1: *We present a convolutional recurrent encoder-decoder network for saliency prediction in driving scenes. Its convolutional structure allows for spatial reasoning, while its recurrent nature is well suited for learning the temporal dependencies of visual behavior in driving scenarios. It takes a sequence of RGB driving frames (first row) as input, and returns a sequence of predicted saliency maps (third row). We train different instances for varied weather and illumination conditions (namely daytime, nighttime, and rain), so that each of them is able to learn the particularities of the conditions. Our results are close to the ground truth (second row), and clearly outperform previous state-of-the-art approaches.*

Abstract

Lately, the automotive industry has experienced a significant development led by the ambitious objective of creating an autonomous vehicle. This entails understanding driving behaviors in different environments, which usually requires gathering and analyzing large amounts of behavioral data from many drivers. However, this is usually a complex and time-consuming task, and data-driven techniques have proven to be a faster, yet robust alternative to modeling drivers' behavior. In this work, we propose a deep learning approach to address this challenging problem. We resort to a novel convolutional recurrent architecture to learn spatio-temporal features of driving behaviors based on RGB sequences of the environment in front of the vehicle. Our model is able to predict drivers' attention in different scenarios while outperforming competing works by a large margin.

CCS Concepts

• ***Computing methodologies*** → ***Interest point and salient region detections;***

1. Introduction

During the last decades, the automotive industry has experienced a significant technological development. To a large extent, this is due to the ambitious objective of creating an autonomous vehicle, in which the driver is relieved of most of their tasks. However, this field is still at an early stage of development, and the driver is still essential for several tasks, most of them involving traffic safety and affecting users both inside and outside of the vehicle.

DOI: 10.2312/ceig.20221149

Currently, and according to recent studies [MHW*17], related-to-attention behaviors (i.e., distraction, fatigue, and aggressive driving) encompass 90% of traffic accidents. This shows how crucial it is for the driver to be alert, to be able to react to unexpected events that may entail a potential risk. To help drivers in their driving tasks, and therefore diminish the risk of traffic accidents, Advanced Driver Assistance Systems (ADAS) are introduced in vehicles. ADAS support drivers by assisting their decision making, and some of them can even partially take driving control over if required. Indeed, ADAS usage has proven to contribute to decreasing the number of accidents that involve drivers' attention [ENW14, Cic17].

Some ADAS are provided with indoor monitoring systems that include head, face, or eye-tracking devices, which capture drivers' visual behavior over time. This data, combined with information from the environment, allow ADAS to detect potentially dangerous maneuvers. However, a part of these indoor devices usually have a prohibitive cost, and are only part of high-segment vehicles. With the surge of data-driven techniques, and motivated by the availability of large datasets of drivers' visual attention [APS*16], some methods have been proposed to predict where drivers look at depending on the visual stimuli they are exposed to while driving [PAS*18], which can be captured with affordable, widespread cameras.

In this work, we present an end-to-end deep-learning approach to predicting drivers' visual attention (see Figure 1). Particularly, and inspired by previous works [NLG19, PAS*18, BLT16], we resort to saliency as a measure of the probability of the driver to direct their attention to each element of an environment, or, in other words, a representation of the interest of each region seen by a driver (see Figures 1 and 2). In this work we propose, for the first time, an encoder-decoder architecture built over the recently presented convolutional long short-term memory networks (ConvLSTM) [SCW*15]. Their recurrent architecture accounts for the temporal dependencies in visual behavior, while their convolutional nature allows the model to learn spatial features (see Section 3).

We train our model over the DR(eye)VE dataset [APS*16], which contains information of visual attention from several drivers in different routes. However, instead of training a single model with the whole dataset, and unlike previous approaches, we split the dataset based on particular illumination and weather conditions (see Figure 1), and train a different instance of the model on each of them. This way, each model can focus on the particularities of each condition, rather than on generalizing to every possible scene. Additionally, to train our models, we augment our data following some image transform operations that resemble common driving circumstances, such as sun flares or shadows, and that have proven to improve the performance and generalization of deep networks [BIK*20]. We finally evaluate our model and compare it to previous approaches for drivers' saliency prediction, with our model outperforming them in most of the scenarios. Moreover, we conduct several ablation studies to endorse the aforementioned design decisions.

Our code is publicly available at https://github.com/DaniMS-ZGZ/DriveRNN.

Figure 2: *Ground truth saliency maps for three different scenes, each depicting a different weather and illumination condition (from top to bottom:* daytime*,* nighttime*, and* rain*). In many cases, drivers' attention is directed to oncoming vehicles or obstacles that could cause a maneuver, anticipating potential dangerous situations.*

2. Related Work

In this section, we briefly review the literature on visual attention prediction, and then focus on the particular case of predicting drivers' attention.

2.1. Visual Attention Prediction

In the last decades, a large body of literature has attempted modeling human visual attention. In 1998, Itti et al. [IKN98] established an original work by computing a saliency map (i.e., a topological representation of the conspicuity of the different elements of an image) from several hand-crafted features. Since then, many works have proposed similar approaches [WK06, ZK11]. Nevertheless, with the surge of data-driven methods, the increase of computational power, and the availability of large datasets [JEDT09, BJB*19], deep-learning based techniques have proven to achieve significantly better results. Within them, convolutional neural networks [VDC14, PSGiN*16, AGiNMO17, KWB16], generative networks [PCM*18, MSB*22], or recurrent neural networks [CBSC18, WSDB18, MGM22], have proven to develop a good ability to extract image features and predict visual attention from them, even in complex scenarios such as virtual reality [SSP*18, MSM20].

2.2. Drivers' Attention Prediction

Lately, many vehicles are being equipped with Advanced Driver Assistance Systems (ADAS), which support the driver in some of their driving tasks, and gather information from both the inside and

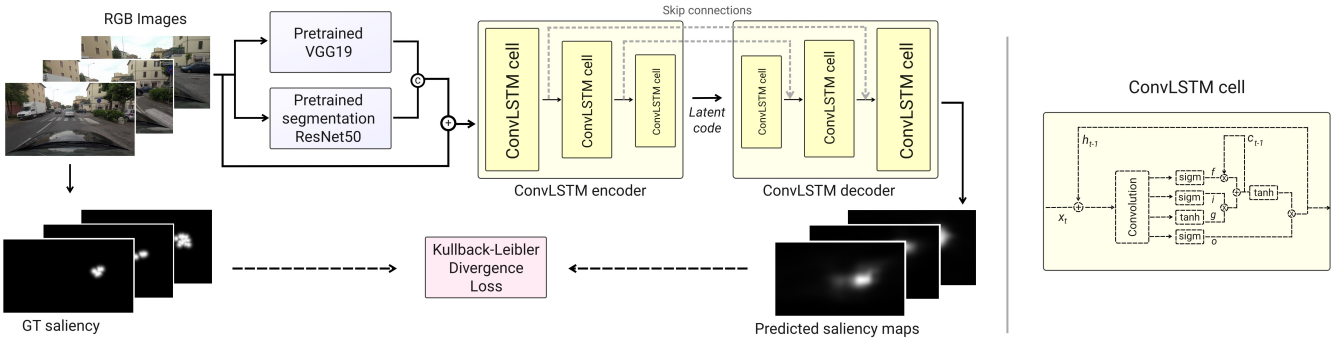

Figure 3: *Overview of the proposed saliency predictor. Our model follows an encoder-decoder architecture, where both modules are composed of three-layer convolutional LSTMs, which are able to extract and parse both the spatial and temporal features of the input image sequence. We feed our model with pairs of sequences of five RGB frames (corresponding to one second of the driving sequence), and their corresponding image latent features extracted with a pretrained VGG19 [RDS*15, SZ14] and a pretrained segmentation network built on ResNet50 [HZRS16]. We include skip connections between both the second and the third layer of both modules, to ease the decoding process. Our model predicts a sequence of the five saliency maps corresponding to the input sequence. Please see Section 3.1 for additional details.*

the outside of the vehicle to assist them during different maneuvers. With the recent development of technology, the increasing affordability of sensors and devices, and the development of artificial intelligence techniques to process large amounts of data, ADAS are evolving towards more and more functionalities, many of them computer-vision-based. Within them, there has been a recent interest on understanding, modeling, and predicting drivers' attention; and saliency has been argued to be essential for an ADAS to anticipate to dangerous situations and better support the driver [APS*16].

Given this, and in order to better understand driving behaviors, some works have been devoted to collecting enough real drivers' data [APS*16]. With them, and given the large body of literature endorsing the use of deep learning for visual attention prediction (see Section 2.1), many works have attempted to leverage all that knowledge to the specific case of drivers' attention prediction. Palazzi et al. [PSC*17,PAS*18] proposed two different approaches for saliency prediction, based on three-dimensional convolutions and a multi-branch approach, while Ning et al. [NLG19] resorted to optical flow as an additional input. In a similar fashion, [FYQX19] proposed a 3D convolutional framework, and included a comprehensive study on the relation between drivers' visual attention and traffic accidents. Aksoy et al. [AYK20] also resorted to CNN to develop a model for predicting when a car should brake depending on saliency information. Other works have also attempted to predict saliency with different techniques, such as semantic augmentation [PMC20] or reinforcement learning [BPK*21]. Different from all of the forementioned works, we resort to a deep recurrent approach to learn and leverage both spatial and temporal features of drivers' attention. Moreover, unlike previous approaches, we propose training different models depending on illumination and weather conditions, rather than a single, generic model. This way, our models are relieved of learning to generalize to every possible situation, and can learn the actual specifics of each condition.

We nevertheless refer the reader to the survey of Kotseruba et al. [KT21] for an exhaustive review on modeling drivers' attention, and the one from Jha et al. [JMH*21] for a large corpus on driving-related datasets.

3. Our Model

We follow a convolutional recurrent approach where, given a sequence of five RGB driving frames, our model is able to predict their corresponding saliency maps. In the following, we provide an in-depth view on our model (Section 3.1) and the loss functions used to train it (Section 3.2). Then, we introduce the dataset that we use to train our model (Section 3.3), and the data augmentation strategy we follow to improve its performance (Section 3.4), together with additional training details (Section 3.5).

3.1. Model Architecture

Our model (see Figure 3) follows an encoder-decoder architecture, where both the encoding and decoding modules are composed of a three-layer convolutional long short-term memory cell (ConvLSTM) [SCW*15], an adaptation of traditional LSTMs to work with convolutional operations. The recurrent structure of ConvLSTMs allows them to extract and leverage temporal dependencies in the data, while their convolutional reformulation endorses their potential to work with spatial information. Specifically, the ConvLSTM used through this work† can be defined as follows:

$$
\begin{aligned}
i_t &= \sigma(Conv(x_t; w_{xi}) + Conv(h_{t-1}; w_{hi}) + b_i) \\
f_t &= \sigma(Conv(x_t; w_{xf}) + Conv(h_{t-1}; w_{hf}) + b_f) \\
o_t &= \sigma(Conv(x_t; w_{xo}) + Conv(h_{t-1}; w_{ho}) + b_o) \\
g_t &= Tanh(Conv(x_t; w_{xg}) + Conv(h_{t-1}; w_{hg}) + b_g) \\
c_t &= f_t \odot c_{t-1} + i_t \odot g_t \\
h_t &= o_t \odot Tanh(c_t),
\end{aligned}
\tag{1}
$$

† https://github.com/ndrplz/ConvLSTM_pytorch

where, for each timestep t, x_t is the input sequence, w_t and b_t are the weights and biases for each of the four convolutional LSTM gates (i, f, o, and g), and $\odot$ is the Hadamard product. Besides, h_t and c_t are the hidden and cell states, as in traditional LSTMs. Please refer to Figure 3, right, for an overview of a ConvLSTM cell.

Besides, and prior to our encoder-decoder block, we include an image feature extractor module [SCW19], composed of a combination of a pretrained VGG19 [RDS*15, SZ14] and a pretrained semantic segmentation network (see Figure 4) built on ResNet50 [HZRS16]. Particularly, we take the last layer of the pretrained VGG19 features (i.e., 512 features), and the 21 output features from the ResNet50. We then resize VGG features to match the size of ResNet50 outputs, and convolve them together with a 1x1 convolution into a single-channel image latent feature [SCW19] that represents the spatial significance of features in the original image, and which is later combined with the RGB sequence to feed our encoder.

We additionally include skip connections to our model. Adding these connections is a common practice in pure convolutional approaches, and we therefore apply them in our problem. This way, our decoder ConvLSTM layers can recover information from their encoder counterparts, thus enhancing their overall decoding ability.

3.2. Loss Function

To train our model, we resort to a loss function based on a weighted combination of mean squared error (MSE) and Kullback-Leibler divergence (KLDiv). On the one hand, MSE performs a pixel-wise evaluation of the error between both predicted and ground-truth saliency maps, offering stability and sensitiveness to outliers. On the other hand, inspired by many state-of-the-art works on saliency prediction [PAS*18, BJO*18], we include a second term based on KLDiv, which measures the difference between a predicted saliency map and its ground truth counterpart as probability distributions.

Thus, our loss function can be defined as follows:

$$\mathcal{L}(P,Q) = \frac{1}{n}\sum_{i=1}^{n}(Q(i)-P(i))^2 + \lambda\sum_{i=1}^{n}P(i)\ln\frac{P(i)}{Q(i)} \qquad (2)$$

where P and Q are the ground truth and predicted saliency maps, and n is the number of pixels in the map. λ regularizes the weight of the last term, and is empirically set to 0.01.

3.3. Dataset

Our model has been trained on the DR(eye)VE dataset [APS*16], which is publicly available, and composed of 74 video sequences of 5 minutes each, recorded at a frame rate of 25 fps (yielding a total of 555,000 frames). These sequences show the vehicle's exterior, towards the front, and have been recorded with a dashboard-mounted camera in a vehicle driven by eight different drivers. The videos were recorded in different surroundings (e.g., cities, highways, secondary roadways), and under varying traffic, light (i.e., day, night) and atmospheric conditions (i.e., sunny, cloudy, rainy).

Figure 4: *Visualization of semantic segmentation for two images corresponding to Düsseldorf (left) and Cologne (right) [COR*16]. Each color represents a different semantic category. Semantic urban scene understanding is critical when driving [COR*16], and we therefore provide our model with semantic information as part of the input.*

Each video sequence is hence composed of 7.500 frames. However, using all of them would require large amounts of memory, and given the frame rate, changes on saliency for consecutive frames are not significant. Therefore, we conduct a frame discretization procedure: For each sequence, we keep one out of every five frames. This way, we reduce each sequence length to 1.500, for a total of 111.000 frames. Since every video is provided with its corresponding saliency map sequence, we follow the same procedure with them. We then split each video in sequences of five frames (i.e., each sequence represents one second) to feed our model. This decision is two-folded: On the one hand, driver's reaction time has been studied to be around 0.9 seconds [ZB07]; while on the other hand too long sequences would hinder LSTMs memory capacity.

Additionally, when curating the dataset, we noticed significant qualitative variations within the RGB images because of the different illumination and weather conditions, which usually translates into drivers having a different behavior (e.g., distractions could be more likely to happen in monotone daytime environments). Therefore, we hypothesize that having a different model for each of these conditions would allow them to learn the particularities of viewing behaviors in them, instead of generalizing for any circumstance. Previous approaches have neglected this fact and trained a single model; unlike them, we decide to divide all the driving sequences into three different scenarios, to then train a different model for each of them:

Daytime conditions (25 driving sequences), during daylight, without rain, in which more brightness is observed. Also, flashes, accentuated shadows, and broader amount of surrounding stimuli (such as pedestrians, cyclists, or other vehicles) can be perceived.

Nighttime conditions (30 driving sequences), from dusk to dawn. Most of the vehicle's environment cannot be fully appreciated, and headlight flares may occur.

Rain conditions (19 driving sequences), during daylight, where the camera lens may be splashed with raindrops and visibility can be partially limited. Outdoor occurrences may also vary due to rain.

However, and in order to perform a fair comparison, we keep a total of nineteen driving sequences for each of the models: fifteen are kept for training and the rest, for testing purposes. Further details on the sequences used in each category can be found in Table 1.

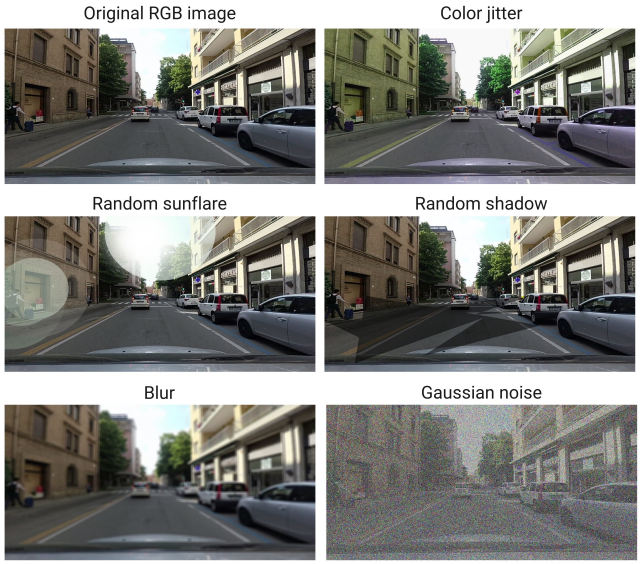

Figure 5: *For a given image (top left), we show the different augmentation techniques we use to train our model. Color jitter (top right) allows our method to be invariant to changes in brightness, contrast, and saturation. Random sunflare (middle left) simulates sun glints in the upper half of the image, whilst random shadow (middle right) simulates low-bright zones in the bottom half of the image. Finally, blur (bottom left) and Gaussian noise (bottom right) introduce noise to our training data and prevent our model from overfitting.*

Although other driving datasets have been presented lately [GRS*21, BPK*21], these have been generated under laboratory conditions, and thus some behaviors could be expected not to be as in real driving scenarios. Thus, we resort to DR(eye)VE as our dataset.

3.4. Data Augmentation

We have observed that, even when grouped (see Section 3.3), the data still has a large variability (e.g., high contrast, lights and shadows accentuation due to sunny scenarios, or cloudy skies that generate more homogeneity in the scene). Such a wide variety may cause a poor adjust of the model to our problem since it is not able to correctly identify these highly changing situations.

To alleviate this, we resort to data augmentation. In particular, we use the *Albumentations* [BIK*20] library, which offers a wide variety of image transformation operations designed to improve deep learning models' overall performance. Particularly, we conduct five different transformations that have been chosen to randomly alter image parameters (see Figure 5), namely color jitter (which randomly changes contrast, brightness, and saturation of the images), random shadows and sunflares (which respectively simulate light glints and shadows that may occur when driving), blur, and Gaussian noise. We set a probability of 30% for each augmentation technique to be applied on each sequence of training data.

Additionally, and to avoid learning the center bias of the data [APS*16], we perform random vertical and horizontal shifting in training images.

Table 1: *We divide the dataset [APS*16] in three subsets, based on different illumination and weather conditions. Each of the subsets contains a total of nineteen sequences, from which fifteen are used for training, and the rest are left for testing purposes.*

Condition	Train sequences ID	Test sequences ID
Daytime	1, 2, 3, 6, 23, 25, 33, 35, 36, 45, 48, 49, 55, 60, 64	11, 34, 40, 42
Nighttime	9, 13, 15, 18, 19, 21, 24, 29, 30, 38, 39, 41, 43, 51, 52	4, 8, 16, 28
Rain	7, 10, 12, 14, 17, 26, 27, 31, 37, 44, 46, 47, 50, 59, 69	5, 32, 63, 74

3.5. Training Details

We implemented our models in PyTorch [PGM*19]. We use the Adam optimization algorithm [KB14], with a learning rate $lr = 10^{-4}$, and we set batch size to 1. We trained our different models on a Nvidia RTX 2080 Ti with 11 GB of VRAM until convergence, for a number of epochs ranging from 30 to 50 depending on the instance of the model, and taking approximately five hours to train each instance.

4. Evaluation

We conduct different qualitative and quantitative evaluations on our results, to assess the performance of our models. In this section, we first introduce the set of metrics used throughout the following studies (Section 4.1) and discuss the results obtained with our different models (Section 4.2). Then, we provide comparisons with existing approaches (Section 4.3), and an exhaustive ablation study to endorse the different design decisions (Section 4.4).

4.1. Metrics

In order to quantitatively evaluate our predicted saliency maps, we resort to three common saliency metrics [BJO*18], namely Pearson's Correlation Coefficient (CC), Kullback-Leibler Divergence (KLDiv), and Normalized Scanpath Saliency (NSS).

Pearson's Correlation Coefficient interprets both saliency maps as random variables, and measures the linear relationship between them as follows:

$$CC(P,Q) = \frac{\sigma(P,Q)}{\sigma(P) \times \sigma(Q)} \tag{3}$$

where P and Q are the predicted and the ground-truth maps, respectively, and $CC(P,Q) \in [-1,1]$, where negative values represent negative correlation, positive values represent correlation, and values close to zero represent uncorrelation.

Normalized Scanpath Saliency (NSS) measures the correspondence between the predicted and the ground truth saliency map as

the average normalized saliency at fixated locations, and is computed as follows:

$$NSS(P,Q) = \frac{1}{N}\sum_{i}^{N}\frac{P-\mu(P)}{\sigma(P)} \times Q(i) \quad (4)$$

where P and Q are the predicted and the ground-truth maps, respectively. Chance is at $NSS(P,Q) = 0$, while higher values indicate better correspondence.

Kullback-Leibler Divergence (KLDiv) measures the difference between two probability distributions. We resort to this metric for optimizing our model, and thus refer to Section 3.2 for further information.

4.2. Results

We have trained three different models, each one for a different driving scenario, mostly motivated by how weather conditions may present different stimuli and impact drivers' attention, namely daytime, nighttime, and rain conditions (see Section 3.3). We evaluate each of these models with a subset of sequences left unseen during training (see Table 1). A quantitative evaluation of these models can be found in Table 2. All three models yield a good performance, with rain conditions being significantly superior to the rest. On the other hand, daytime conditions yield a slightly lower performance. We hypothesize that this is due to our data division, since daytime conditions also usually present more varied contrast and brightness changes, and drivers are more likely to be distracted, which may be hindering the ability of that network to generalize.

Qualitative results of our model can be seen in Figure 6. We show different predictions on scenarios from all three illumination and weather conditions in sets of three rows, showing, from top to bottom, the original RGB image, the ground truth (GT) saliency map, and the predicted (Pred.) saliency map. Besides, each RGB frame includes its corresponding weather condition overlaid in the top-right. Our predicted saliency maps resemble the ground truth ones, focusing on relevant areas, such as moving vehicles (first row, first column), or streets where the vehicle is turning towards (third row, fourth column).

4.3. Comparisons

We have also compared our models to previous approaches. Although other driver attention prediction approaches have been presented lately [BPK*21, GRS*21], these works have been trained or tested on other datasets that have been generated under laboratory conditions (see Section 3.3), and we thus compare ourselves only to those that trained and tested their model on the DR(eye)VE dataset [MS14, WSS15, WSP15, CBSC16, BLT16, PSC*17, PAS*18, NLG19], as we did throughout this work.

Table 2 shows the results of the conducted quantitative comparisons. First and second rows are two lower baselines included for completeness: Gaussian baseline represents a centered heat map, while mean baseline is the average of all training fixation maps [PAS*18]. Third to tenth rows include the different aforementioned saliency prediction approaches, while last three rows show the performance of each of our specific models, which outperform previous approaches in most of the scenarios.

Table 2: *We have compared our model to previous approaches for saliency prediction using two common saliency metrics, namely CC and KLDiv (see Section 4.1). Here, we include two lower baselines, namely a Gaussian baseline (representing a centered heat map), and a mean baseline (i.e., the average of all training fixations), and eight existing saliency prediction approaches. The last three rows show the values yielded by our specific models. Arrows show whether higher or lower is better, and best results are boldfaced. Our proposed specific models outperform previous approaches in most of the scenarios. Please refer to Section 4.3 for further discussion.*

Model	CC↑	KLDiv↓
Baseline Gaussian	0.40	2.16
Baseline Mean	0.51	1.60
Mathe et al. [MS14]	0.04	3.30
Wang et al. [WSP15]	0.04	3.40
Wang et al. [WSS15]	0.11	3.06
MLNet [CBSC16]	0.44	2.00
RMDN [BLT16]	0.41	1.77
Palazzi et al. [PSC*17]	0.55	1.48
Palazzi et al. [PAS*18]	0.56	**1.40**
Ning et al. [NLG19]	0.57	1.50
Ours (daytime)	**0.59**	1.46
Ours (nighttime)	**0.61**	**1.40**
Ours (rain)	**0.70**	**1.06**

4.4. Ablation Studies

In order to endorse the different design decisions on our model, we conduct several ablation studies. We resort to the metrics introduced in Section 4.1 to quantitatively evaluate those decisions. Results can be seen in Table 3.

Model architecture. Previous works (see Section 4.3) followed convolutional approaches. Taking this into consideration, we developed an encoder-decoder architecture similar to our proposed recurrent one (i.e., three encoding and decoding layers, skip connections, and semantic segmentation), where both modules consisted of convolutional layers instead of ConvLSTM layers. Our proposed ConvLSTM model (last row) significantly outperforms any convolutional approach (first three rows), since the latter are only provided with spatial features, while the former is able to process, extract, and leverage both spatial and temporal features, which are crucial for modeling human visual attention.

Data Augmentation. During the design process, we also analyzed the use of data augmentation while training the model (see Section 3.4). Data augmentation has proven to be very effective to improve performance and reduce overfitting [BIK*20]. Indeed, when we carry out our data augmentation strategy (third row), results are significantly better than when using no augmentation technique (second row), verifying the benefits of this kind of procedures.

Illumination and weather conditions. Finally, we evaluate

Figure 6: *Qualitative results of our model. We show different saliency predictions in sets of three rows, where the first one depicts the original RGB frame, the second one corresponds to the ground truth (GT) saliency map, and the third one is the predicted (Pred.) saliency map. We include whether the image is from daytime, nighttime, or rain conditions (see Section 3.3) overlaid on the top-right of each RGB frame. Our predicted saliency maps closely resemble the ground truth ones, focusing on relevant parts, such as moving vehicles (first row, first column), or the streets the car is turning towards (third row, fourth column). Please refer to Table 2 for a quantitative evaluation.*

Table 3: *Results of our ablation studies. We have evaluated the effectiveness of our data augmentation strategy (second and third rows, see Section 3.4 for additional details), which generates a consistent improvement in performance. We have also analyzed the benefits of training a specific model for a particular illumination and weather condition instead of training a single model with the whole training set (first and second rows, see Section 3.3 for additional details), obtaining results that ratify the usage of several, more specialized models. Additionally, given the temporal component of visual attention while driving, we have evaluated whether a recurrent neural network is better suited than a convolutional model (third and fourth row, see Section 3.1), since results prove that the ability of RNNs to handle temporal dependencies significantly improves the overall performance of the model. Please refer to Section 4.4 for further discussion on these studies.*

Model	Condition	CC↑	KLDiv↓	NSS↑
Generic conv. model *(no augmentation)*	-	0.41	2.25	3.45
Weather conv. model *(no augmentation)*	Daytime	0.37	2.37	3.24
	Nighttime	0.44	2.07	3.47
	Rain	0.47	1.98	4.03
Weather conv. model *(w/ augmentation)*	Daytime	0.45	2.19	3.88
	Nighttime	0.53	1.82	4.22
	Rain	0.57	1.73	5.20
Generic ConvLSTM model *(w/ augmentation)* ***(ours)***	-	0.60	1.43	4.76
Weather ConvLSTM model *(w/ augmentation)* ***(ours)***	Daytime	0.59	1.46	4.37
	Nighttime	0.61	1.40	4.25
	Rain	**0.70**	**1.06**	**5.74**

whether dividing the dataset into three different illumination and weather conditions can lead to better results. We compare a generic model trained following the same train-test data separation as in Palazzi et al. [PAS*18] to our three specific models, both in an only-convolutional (first row compared to second row) and recurrent (fourth row compared to fifth row) fashion, trained as commented in Section 3.3. With our proposed division, nighttime and rain models' performance increase, achieving more accurate predictions. However, it is noteworthy that daytime conditions slightly worsen. This behavior holds in every tested model configuration, and we believe this is due to the great variability that exists in the daytime pictures. However, overall results successfully prove separating the data into different categories, which is also supported by the fact that drivers' behavior changes depending on the environment, as well as the insufficient ability of a single model to learn and generalize attentional patterns in many different situations.

5. Conclusion

In conclusion, we have proposed a convolutional recurrent approach for saliency prediction in driving image sequences. We leverage a combination of a pretrained VGG19 and a segmentation ResNet50 network to generate image latent features, and resort to a convolutional recurrent encoder-decoder architecture with skip connections to extract spatio-temporal features and predict saliency from them.

We have trained three different models, each of them devoted to a different subset of possible driving scenarios in terms of illumination and weather conditions (i.e., daytime, nighttime, and rain conditions), and which outperform previous state-of-the-art methods. Additionally, we have conducted several ablation studies to endorse the different design decisions on our model.

5.1. Limitations and Future Work

The proposed approach predicts saliency based on a sequence of RGB images. However, information about the driver's psychological and physical parameters (see, e.g., [CMPR15]) can be of great relevance to anticipate drivers' maneuvers. To this end, providing our network with data from different sensors [UKT94] remains an interesting avenue for future work. Besides additional information of the driver, our model could also benefit from having additional information from the environment, either as additional visual information (such as lidar, radar, etc.), or as multimodal cues. Specifically, auditory cues (such as ambulance sirens or another vehicle's horn) may influence driver's attention.

Generating three different models adapted to light and weather conditions involves selecting the appropriate model for each situation. In this work, such classification has been carried out manually, considering the light conditions (i.e., daytime and nighttime) first, and then classifying daytime images based on further weather conditions (i.e., rainy), as explained in Subsection 3.3. However, in order to process unlabelled data, our model would require either a manual selection or a classification method to decide the most suitable model for each case.

We believe that having a model able to accurately and rapidly predict driver's attention could have a significant impact on many applications, including developing driving simulators to train elder drivers' visual attention [HBMH18]; understanding human behavior while doing particular driving tasks [FPBBT13, HWC06], or even enhancing current ADAS by increasing traffic safety, for instance by ensuring that drivers' attention is directed towards relevant or dangerous stimuli. We believe our work is a timely effort and an important step towards modeling and understanding driving behaviors.

Acknowledgments

This work has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (project CHAMELEON, Grant No 682080), and the Marie Skłodowska-Curie grant agreement No 956585. This project was also supported by a 2020 Leonardo Grant for Researchers and Cultural Creators, BBVA Foundation (the BBVA Foundation accepts no responsibility for the opinions, statements and contents included in the project and/or the results thereof, which are entirely the responsibility of the authors). This work has also received funding from Spain's Agencia Estatal de Investigación (project PID2019-105004GB-I00) and Gobierno de Aragon (T34_20R). Additionally, Daniel Martin was supported by a Gobierno de Aragon (2020-2024) predoctoral grant.

References

[AGiNMO17] ASSENS M., GIRO-I NIETO X., MCGUINNESS K., O'CONNOR N. E.: Saltinet: Scan-path prediction on 360 degree images using saliency volumes. In *Proceedings of the IEEE International Conference on Computer Vision Workshops* (2017), pp. 2331–2338. 2

[APS*16] ALLETTO S., PALAZZI A., SOLERA F., CALDERARA S., CUCCHIARA R.: Dr (eye) ve: a dataset for attention-based tasks with applications to autonomous and assisted driving. In *Proceedings of the ieee conference on computer vision and pattern recognition workshops* (2016), pp. 54–60. 2, 3, 4, 5

[AYK20] AKSOY E., YAZICI A., KASAP M.: See, attend and brake: An attention-based saliency map prediction model for end-to-end driving. *arXiv preprint arXiv:2002.11020* (2020). 3

[BIK*20] BUSLAEV A., IGLOVIKOV V. I., KHVEDCHENYA E., PARINOV A., DRUZHININ M., KALININ A. A.: Albumentations: Fast and flexible image augmentations. *Information 11*, 2 (2020). 2, 5, 6

[BJB*19] BYLINSKII Z., JUDD T., BORJI A., ITTI L., DURAND F., OLIVA A., TORRALBA A.: Mit saliency benchmark. http://saliency.mit.edu/, 2019. 2

[BJO*18] BYLINSKII Z., JUDD T., OLIVA A., TORRALBA A., DURAND F.: What do different evaluation metrics tell us about saliency models? *IEEE transactions on pattern analysis and machine intelligence 41*, 3 (2018), 740–757. 4, 5

[BLT16] BAZZANI L., LAROCHELLE H., TORRESANI L.: Recurrent mixture density network for spatiotemporal visual attention. *arXiv preprint arXiv:1603.08199* (2016). 2, 6

[BPK*21] BAEE S., PAKDAMANIAN E., KIM I., FENG L., ORDONEZ V., BARNES L.: Medirl: Predicting the visual attention of drivers via maximum entropy deep inverse reinforcement learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 13178–13188. 3, 5, 6

[CBSC16] CORNIA M., BARALDI L., SERRA G., CUCCHIARA R.: A deep multi-level network for saliency prediction. In *2016 23rd International Conference on Pattern Recognition (ICPR)* (2016), IEEE, pp. 3488–3493. 6

[CBSC18] CORNIA M., BARALDI L., SERRA G., CUCCHIARA R.: Predicting human eye fixations via an lstm-based saliency attentive model. *IEEE Transactions on Image Processing 27*, 10 (2018), 5142–5154. 2

[Cic17] CICCHINO J. B.: Effectiveness of forward collision warning and autonomous emergency braking systems in reducing front-to-rear crash rates. *Accident Analysis & Prevention 99* (2017), 142–152. 2

[CMPR15] CHACON-MURGUIA M. I., PRIETO-RESENDIZ C.: Detecting driver drowsiness: A survey of system designs and technology. *IEEE Consumer Electronics Magazine 4*, 4 (2015), 107–119. 8

[COR*16] CORDTS M., OMRAN M., RAMOS S., REHFELD T., ENZWEILER M., BENENSON R., FRANKE U., ROTH S., SCHIELE B.: The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 3213–3223. 4

[ENW14] EDWARDS M., NATHANSON A., WISCH M.: Estimate of potential benefit for europe of fitting autonomous emergency braking (aeb) systems for pedestrian protection to passenger cars. *Traffic injury prevention 15*, sup1 (2014), S173–S182. 2

[FPBBT13] FREYDIER C., PAXION J., BERTHELON C., BASTIEN-TONIAZZO M.: Divided-attention task on driving simulator: comparison among three groups of drivers. In *11th International Conference on Naturalistic Decision Making 2013* (2013), ARPEGE SCIENCE PUBLISHING, p. 4p. 8

[FYQX19] FANG J., YAN D., QIAO J., XUE J.: Dada: A large-scale benchmark and model for driver attention prediction in accidental scenarios. *arXiv preprint arXiv:1912.12148* (2019). 3

[GRS*21] GOPINATH D., ROSMAN G., STENT S., TERAHATA K., FLETCHER L., ARGALL B., LEONARD J.: Maad: A model and dataset for" attended awareness" in driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 3426–3436. 5, 6

[HBMH18] HAEGER M., BOCK O., MEMMERT D., HÜTTERMANN S.: Can driving-simulator training enhance visual attention, cognition, and physical functioning in older adults? *Journal of aging research 2018* (2018). 8

[HWC06] HORREY W. J., WICKENS C. D., CONSALUS K. P.: Modeling drivers' visual attention allocation while interacting with in-vehicle technologies. *Journal of Experimental Psychology: Applied 12*, 2 (2006), 67. 8

[HZRS16] HE K., ZHANG X., REN S., SUN J.: Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 770–778. 3, 4

[IKN98] ITTI L., KOCH C., NIEBUR E.: A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence 20*, 11 (1998), 1254–1259. 2

[JEDT09] JUDD T., EHINGER K., DURAND F., TORRALBA A.: Learning to predict where humans look. In *IEEE ICCV* (2009), IEEE, pp. 2106–2113. 2

[JMH*21] JHA S., MARZBAN M. F., HU T., MAHMOUD M. H., AL-DHAHIR N., BUSSO C.: The multimodal driver monitoring database: A naturalistic corpus to study driver attention. *IEEE Transactions on Intelligent Transportation Systems* (2021). 3

[KB14] KINGMA D. P., BA J.: Adam: A method for stochastic optimization. In *ICLR* (2014). Last updated in arXiv in 2017. 5

[KT21] KOTSERUBA I., TSOTSOS J. K.: Behavioral research and practical models of drivers' attention. *arXiv preprint arXiv:2104.05677* (2021). 3

[KWB16] KÜMMERER M., WALLIS T. S. A., BETHGE M.: Deepgaze ii: Reading fixations from deep features trained on object recognition. *arXiv preprint arXiv:1610.01563* (2016). 2

[MGM22] MARTIN D., GUTIERREZ D., MASIA B.: A probabilistic time-evolving approach to scanpath prediction. In *arxiv.2204.09404* (2022), arXiv. 2

[MHW*17] MARTINEZ C. M., HEUCKE M., WANG F.-Y., GAO B., CAO D.: Driving style recognition for intelligent vehicle control and advanced driver assistance: A survey. *IEEE Transactions on Intelligent Transportation Systems 19*, 3 (2017), 666–676. 2

[MS14] MATHE S., SMINCHISESCU C.: Actions in the eye: Dynamic gaze datasets and learnt saliency models for visual recognition. *IEEE transactions on pattern analysis and machine intelligence 37*, 7 (2014), 1408–1424. 6

[MSB*22] MARTIN D., SERRANO A., BERGMAN A. W., WETZSTEIN G., MASIA B.: Scangan360: A generative model of realistic scanpaths for 360 images. *IEEE Transactions on Visualization & Computer Graphics*, 01 (2022), 1–1. 2

[MSM20] MARTIN D., SERRANO A., MASIA B.: Panoramic convolutions for 360° single-image saliency prediction. In *CVPR Workshop on Computer Vision for Augmented and Virtual Reality* (2020). 2

[NLG19] NING M., LU C., GONG J.: An efficient model for driving focus of attention prediction using deep learning. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)* (2019), IEEE, pp. 1192–1197. 2, 3, 6

[PAS*18] PALAZZI A., ABATI D., SOLERA F., CUCCHIARA R., ET AL.: Predicting the driver's focus of attention: the dr (eye) ve project. *IEEE transactions on pattern analysis and machine intelligence 41*, 7 (2018), 1720–1733. 2, 3, 4, 6, 8

[PCM*18] PAN J., CANTON C., MCGUINNESS K., O'CONNOR N. E., TORRES J., SAYROL E., GIRO-I NIETO X. A.: Salgan: Visual saliency prediction with generative adversarial networks. *arXiv preprint arXiv:1701.01081* (2018). 2

[PGM*19] PASZKE A., GROSS S., MASSA F., LERER A., BRADBURY J., CHANAN G., KILLEEN T., LIN Z., GIMELSHEIN N., ANTIGA L., ET AL.: Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems 32* (2019). 5

[PMC20] PAL A., MONDAL S., CHRISTENSEN H. I.: " looking at the right stuff"-guided semantic-gaze for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 11883–11892. 3

[PSC*17] PALAZZI A., SOLERA F., CALDERARA S., ALLETTO S., CUCCHIARA R.: Learning where to attend like a human driver. In *2017 IEEE Intelligent Vehicles Symposium (IV)* (2017), IEEE, pp. 920–925. 3, 6

[PSGiN*16] PAN J., SAYROL E., GIRO-I NIETO X., MCGUINNESS K., O'CONNOR N. E.: Shallow and deep convolutional networks for saliency prediction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2016). 2

[RDS*15] RUSSAKOVSKY O., DENG J., SU H., KRAUSE J., SATHEESH S., MA S., HUANG Z., KARPATHY A., KHOSLA A., BERNSTEIN M., ET AL.: Imagenet large scale visual recognition challenge. *International journal of computer vision 115*, 3 (2015), 211–252. 3, 4

[SCW*15] SHI X., CHEN Z., WANG H., YEUNG D.-Y., WONG W.-K., WOO W.-C.: Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1* (Cambridge, MA, USA, 2015), NIPS'15, MIT Press, p. 802–810. 2, 3

[SCW19] SUN W., CHEN Z., WU F.: Visual scanpath prediction using ior-roi recurrent mixture density network. *IEEE Transactions on Pattern Analysis and Machine Intelligence 43*, 6 (2019), 2101–2118. 4

[SSP*18] SITZMANN V., SERRANO A., PAVEL A., AGRAWALA M., GUTIERREZ D., MASIA B., WETZSTEIN G.: Saliency in vr: How do people explore virtual environments? *IEEE Transactions on Visualization and Computer Graphics 36*, 4 (2018). 2

[SZ14] SIMONYAN K., ZISSERMAN A.: Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014). 3, 4

[UKT94] UENO H., KANEDA M., TSUKINO M.: Development of drowsiness detection system. In *Proceedings of VNIS'94-1994 Vehicle Navigation and Information Systems Conference* (1994), IEEE, pp. 15–20. 8

[VDC14] VIG E., DORR M., COX D.: Large-scale optimization of hierarchical features for saliency prediction in natural images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2014). 2

[WK06] WALTHER D., KOCH C.: Modeling attention to salient proto-objects. *Neural Networks 19* (2006), 1395–1407. 2

[WSDB18] WANG W., SHEN J., DONG X., BORJI A.: Salient object detection driven by fixation prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2018). 2

[WSP15] WANG W., SHEN J., PORIKLI F.: Saliency-aware geodesic video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2015), pp. 3395–3402. 6

[WSS15] WANG W., SHEN J., SHAO L.: Consistent video saliency using local gradient flow optimization and global refinement. *IEEE Transactions on Image Processing 24*, 11 (2015), 4185–4196. 6

[ZB07] ZHANG X., BHAM G. H.: Estimation of driver reaction time from detailed vehicle trajectory data. *Moas 7* (2007), 574–579. 4

[ZK11] ZHAO Q., KOCH C.: Learning a saliency map using fixated locations in natural scenes. *Journal of Vision 11* (2011), 9. 2